Rec'd PCT/PTC 14 APR 2005 10/531020

PCT/SE 03 / 0 0 0 2 2

21-01-2003

REC'D 28 JAN 2003

WIPO PCT

PRV

PATENT- OCH REGISTRERINGSVERKET
Patentavdelningen

Intyg
Certificate

PATENT- OCH REGISTRERINGSVERKET

Härmed intygas att bifogade kopior överensstämmer med de handlingar som ursprungligen ingivits till Patent- och registreringsverket i nedannämnda ansökan.

This is to certify that the annexed is a true copy of the documents as originally filed with the Patent- and Registration Office in connection with the following patent application.

(71) Sökande Applicant (s) *Telefonaktiebolaget L M Ericsson (publ), Stockholm SE*

(21) Patentansökningsnummer Patent application number *0203104-5*

(86) Ingivningsdatum Date of filing *2002-10-18*

Stockholm, 2003-01-13

För Patent- och registreringsverket
For the Patent- and Registration Office

Lina Oljeqvist

Lina Oljeqvist

Avgift
Fee

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

BEST AVAILABLE COPY

+46 8 7520937

Uppgjord (även faktaansvarig om annan)	Nr - No.		
EAB/LA/B J Petersson / H Emilion	EAB/QSI -2002:073 Uen		
Dokansv/Godk - Doc respons/Approved / Kontr - Checked	Datum - Date	Rev	File
EAB/QSIC (K. Boestad)	10/18/02	A	

Method and Apparatus for Network-Initiated Rate Control for Person-to-Content Services in a Mobile System

Names of the Inventors
Svetlana Chemiakina
Luigi D'Antonio
Justus Petersson
Robert Skog

1 TECHNICAL FIELD

End-user quality of service optimisation for packet switched services in a mobile network, such as WCDMA, CDMA2000 and GPRS.

The invention is applicable to person-to-content packet switched services in a mobile system.

2 TECHNICAL BACKGROUND

2.1 THE PROBLEM AREA

The invention makes use of the assigned/employed bandwidth for a particular session. This bandwidth information is retrieved from a radio access network and facilitates enhanced quality-of-service for mobile network agnostic applications.

2.2 STATE OF THE ART

There exist numerous solutions on how to improve the quality of service (QoS) for an end-user who is residing in a mobile system.

In EP1126716, the concept of using radio network information from the mobile network in order to faster and more accurately regulate the application bit rate for video data services is introduced.

Except for the patent application mentioned above, previous works in the field of enhancing the QoS for video/streaming services in wireless networks have not been focused on the idea of taking the present radio resource information into account as a means for regulating the application bit rate.

In [G. Côté, S. Shirani and F. Kossentini, "Optimal Mode Selection and Synchronisation for Robust Video Communications over Error-Prone Networks", IEEE Journal on Selected Areas in Communications, vol.18, No.6, June 2000] and [C. Hsu, A. Ortega and M. Khansari, "Rate Control for Robust Video Transmission over Burst-Error Wireless Channels", IEEE Journal on Selected Areas in Communications, vol.17, No.5, May 1999], methods on how the channel state information may be utilized for making a more precise judgement for server controlled bit-rate regulation are discussed. In [G. Cheung and T. Yoshimura, "Streaming Agent: A Network Proxy for Media Streaming in 3G Wireless Networks", IEEE Packet Video Workshop 2002], a Streaming Agent (SA) between the wired and wireless network

Read and Understood by:.. Date:............

Read and Understood by:.. Date:............

Uppgjord (även faktaansvarig om annan) - Prepared (also subject responsible if other)		Nr - No.		
EAB/LA/B J Petersson / H Emilion		EAB/QSI -2002:073 Uen		
Dokansv/Godk - Doc respons/Approved	Kontr - Checked	Datum - Date	Rev	File
EAB/QSIC (K. Boestad)		10/18/02	A	

was introduced in order to minimize the impacts of congestion situations in the wired link under the assumption that the wireless link was almost loss-less. The SA sends timely feedback to the Server to track the wired link state. Thus, the Server may take more appropriate rate-switch decisions. In [A. Schieder et al., "Resource Efficient Streaming Bearer Concept for GERAN", WPMC October 2002 in Honolulu] is disclosed a method on how the buffer fullness in the Client may be used as a trigger for feedback messages to the Server. The goal was to efficiently utilize the radio resources in a GSM/EDGE radio access network and at the same time provide the end-user with good QoS.

2.3 PROBLEMS

The ideas mentioned above try to solve the QoS optimisation problem for the end-user in a mobile network primarily for Streaming/Video services.

Eg. in EP1126716, a solution is presented which is targeted for UDP based services. That is, the optimisation issue for TCP-based applications is not addressed. However, this invention is partly based on that patent application.

In [G. Cheung and T. Yoshimura, "Streaming Agent: A Network Proxy for Media Streaming in 3G Wireless Networks", IEEE Packet Video Workshop 2002], the writers assume that the bottleneck is in the wired link and not in the wireless link and propose an idea on how to optimise this problem for UDP based services.

The rest of the prior art found has a clear client-centric approach, that is, it does not involve the information from the radio access network. These inventions, too, are aimed at UDP based services.

3 THE INVENTION

3.1 SUMMARY

The main principles of the idea are the following:

Information about the assigned/employed bandwidth over the air-interface used for the session is sent from a radio network controlling (RNC) entity to an intermediate node, a so called Service Network Session Database (SNSD). The SNSD is connected to a QoS-Proxy. The QoS-Proxy has the capabilities of optimising the end-user's quality of service on the basis of the employed bandwidth information, which it receives from the RNC entity and its own internal algorithms. (How these algorithms work are not a subject for this invention).

3.2 DESCRIPTION

The figure below depicts the logic view of the solution.

Read and Understood by:.. Date:............

Read and Understood by:.. Date:............

Uppgjord - Prepared	Nr - No.
EAB/LA/B J Petersson / H Emilion	EAB/QSI -2002:073 Uen
Dokansv/Godk - Doc respons/Approved: EAB/QSIQ (K. Bogstad)	Datum - Date: 10/18/02 — Rev: A — File

2002-10-18

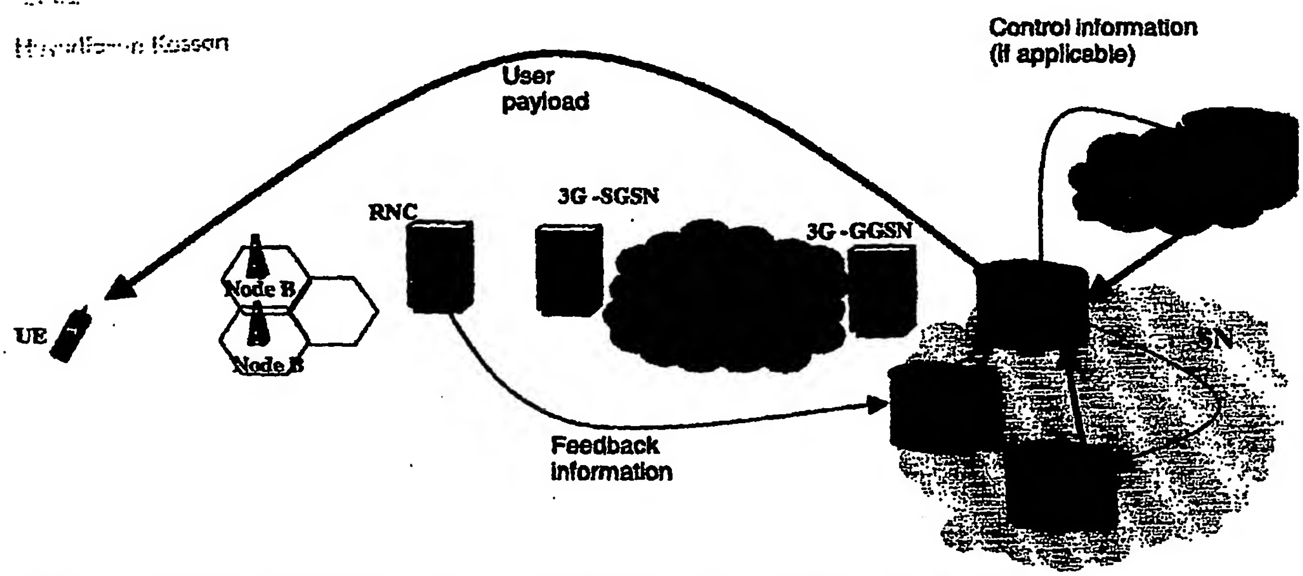

Figure 1 The SNSD receives rate control information from a Radio Network Controling node, e.g. an RNC or a BSC in a UMTS/GSM system. This information may be used for optimising UDP and TCP based services as well as trigger for possible filtering functionalities in the QoS-Proxy.

3.3 INITIALISATION AND RATE CONTROL NOTIFICATION

The RNC needs to be notified whether or not a session requires the feedback service. This notification may be accomplished during the i) configuration or ii) the session set-up.

One way to convey the rate information to the SNSD is to use the established GTP-U tunnel between the RNC and the GGSN. This, however, requires extra functionalities in the GGSN. Another approach may be to directly inform the RNC about the SNSD's IP address upon i) configuration or ii) session set-up. (By doing so, the RNC updates the SNSD about valid bit rate without any interrogation from an intermediate node).

3.4 SNSD FUNCTIONALITY

The SNSD retrieves radio network feedback information (rate control information) from a Radio Network Controlling node (e.g. an RNC or a BSC). This information is either temporarily stored or directly relayed via the QoS-Proxy to an Application Server.

3.5 RATE CONTROL ADAPTATION

3.5.1 Pre-requisities of rate control adaptation

The user equipment (UE) has activated a packet data protocol (PDP) context. This PDP context includes among other information an access point name (APN). That

Read and Understood by:.. Date:..............

Read and Understood by:.. Date:..............

Uppgjord (även faktaansvarig om annan) - Prepared (also subject responsible if other)		Nr - No.		
EAB/LA/B J Petersson / H Emilion		EAB/QSI -2002:073 Uen		
Dokansv/Godk - Doc respons/Approved	Kontr - Checked	Datum - Date	Rev	File
EAB/QSIC (K. Boestad)		10/18/02	A	

APN gives a logical connection between UE and SNSD. In the SNSD the following information may be stored per UE :
1) IP-address of the UE
2) Bit-rate
3) plus other information, eg. user's MSISDN, not related to this invention.

3.5.2 Downloading or Web Surfing (on top of the TCP protocol)

The characteristics of the air-interface, e.g. fading dips and shadowing, may have negative consequences for the end-user. This is especially true for applications that use the TCP protocol as a transport bearer. Eg. a long latency over the air-interface may trigger the TCP congestion avoidance mechanism, leading to less bandwidth for the session and resulting in a very lousy performance for the end-user. On the other hand, if the end-user will temporarily get increased bandwidth over the air-interface, this may most likely not speed up the TCP connection to the same extent, implying that scarce radio resources will not be utilised.

Since the SNSD interworks (provides) the QoS-Proxy with the currently employed bit rate over the air-interface, the QoS-Proxy has the ability to set the TCP-parameters such as eg. the Segment/Window sizes to optimally fit the radio resource situation.

The flow diagram in Figure 2 below gives an example on how one may optimise the end-user's QoS for TCP based services (e.g., for downloading or web surfing).

Read and Understood by:.. Date:............

Read and Understood by:.. Date:............

Uppgjord (även faktaansvarig om annan) - Prepared (also subject responsible if other)		Nr - No.		
EAB/LA/B J Petersson / H Emilion		EAB/QSI -2002:073 Uen		
Dokansv/Godk - Doc respons/Approved	Kontr - Checked	Datum - Date	Rev	File
EAB/QSIC (K. Boestad)		10/18/02	A	

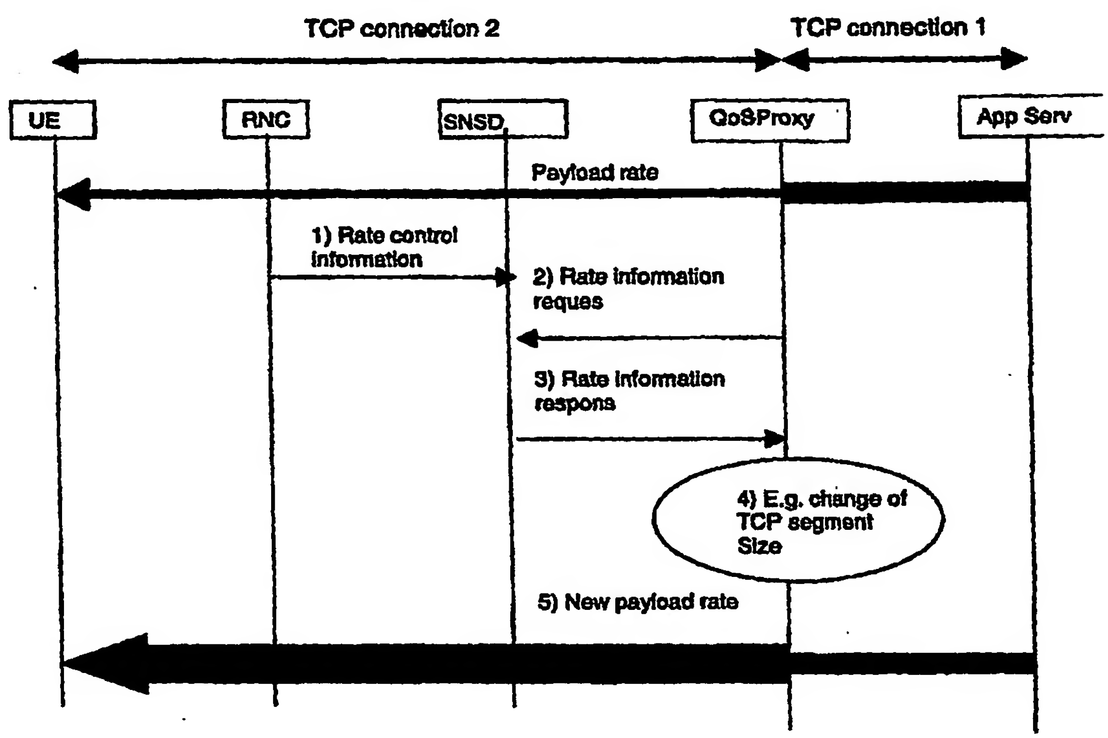

Figure 2 Adaptive TCP optimisation according to the limited resources controlled by the RNC. In order to handle TCP adaptation optimally, we have split the connection into two parts, TCP connection 1 and 2, respectively. Note that this picture is a "snapshot" of the "Rate Information" service behaviour. That is, for one whole session the above "algorithm" may be employed many times.

Initial conditions for the illustrated example are these:

The Application Server sends the payload at a certain bit rate according to the TCP mechanisms specified by IETF. Due to limitations over the air-interface the QoS-Proxy is not allowed to relay the incoming payload at the same pace as it arrives. Therefore, the QoS-Proxy temporarily stores the incoming payloads in a cache (not shown in the figure) and acknowledges the Application Server (by sending "ACKs"), as if they were sent to the Client. By doing so one decreases the risk that the TCP congestion control mechanism will be employed. At the same time one minimises the total download time for the requested object/file.

Alternatively, the TCP connection can go directly from the UE to the Application Server, likewise the rate control information can go directly from the RNC entity to the Application server.

Read and Understood by:.. Date:............

Read and Understood by:.. Date:............

Uppgjord (även faktaansvarig om annan) - +46 8 7520937		Nr - No.		
EAB/LA/B J Petersson / H Emilion		EAB/QSI -2002:073 Uen		
Dokansv/Godk - Doc respons/Approved	Kontr - Checked	Datum - Date	Rev	File
EAB/QSIC (K. Boestad)		10/18/02	A	

1) The Radio Network Controlling entity, e.g. the RNC, has discovered that the system has spare capacity and thus informs the SNSD that a specific session may enjoy a new and higher bit rate.
2) The QoS-Proxy may periodically check the SNSD about the currently allowed transport bit rate for a given session.
3) In the above case the QoS-Proxy gets a response from the SNSD saying that the RNC allows for a higher bit rate for a given session.
4) The QoS-Proxy's internal wireless TCP optimisation algoritms adapt to the new situation.
Note that the bit rate out from the QoS-Proxy may temporarily by greater than the incoming bit rate, since it has "old" payload in its cache.
5) The end-user receives the content at the new bit rate.

3.5.3 Content transformation

Imagine that a user is in the beginning of downloading a large picture and that the cell he is residing in is currently congested. Probably the user will get annoyed with the long latency it takes for the download and will thus disconnect the session. An alternative to this may be to speed up the downloading process by filtering out some information in the picture. Of course, the rendered quality of the picture will be decreased compared to the whole information being transmitted. However, it can be argued that the latency may have more severe impact on the end-users' total quality-of-service experience than anything else.

In Figure 3 an example is shown that visualizes this.

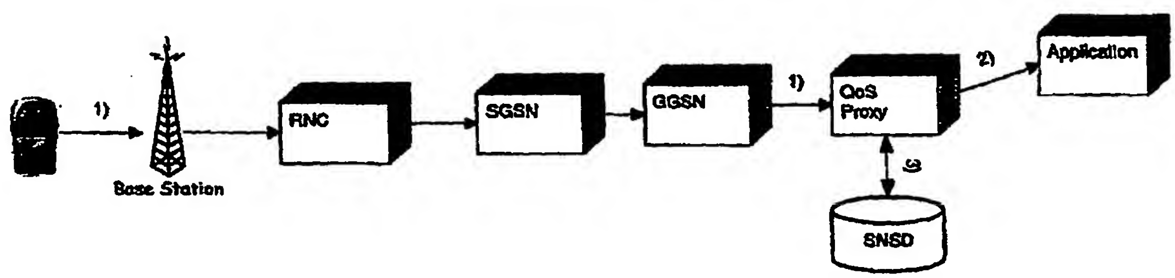

Figure 3

1) The UE requests a content from the Application Server (which could be eg. a MMS Server). The request is directly or indirectly captured by the QoS Proxy.

2) QoS Proxy fetches the content from the Application Server.

3) Before sending the content back to UE the QoS Proxy looks in the SNSD to find out what is current available bit rate for the UE.

   IP-address can be used to find session information about the UE in SNSD.

Read and Understood by:.. Date:.............

Read and Understood by:.. Date:.............

Uppgjord (även faktaansvarig om annan) - Prepared (also subject responsible if other)		Nr - No.		
EAB/LA/B J Petersson / H Emilion		EAB/QSI -2002:073 Uen		
Dokansv/Godk - Doc respons/Approved	Kontr - Checked	Datum - Date	Rev	File
EAB/QSIC (K. Boestad)		10/18/02	A	

Current bit rate is received. Depending on some QoS algorithm (where current bit rate is one parameter) some transformation might be needed for the content before delivery to UE. An example of such transformation can be reducing the quality of picture in order to reduce the transmission time.

3.5.4 Streaming (on top of the UDP protocol)

When the UDP protocol is employed one has greater opportunity to utilise the radio resources compared to using TCP. (The UDP protocol of today has no congestion avoidance mechanisms). The content residing close to the Streaming server is often coded in different bit rates. This may be utilised in a mobile environment.

Imagine that a user has set up a stream with the bit rate 2X bits/s and that the systems discover a congestion/decongestion situation. This information may quickly be relayed to the SNSD/QoS-Proxy, which may trigger the originating source (ie, the Streaming Server) to perform a down/up-switch to 1X or 3X bits/s, respectively.

The flow diagram in Figure 3 below gives an example how one may optimise the end-user's QoS experience for the UDP based services (e.g., for streaming services).

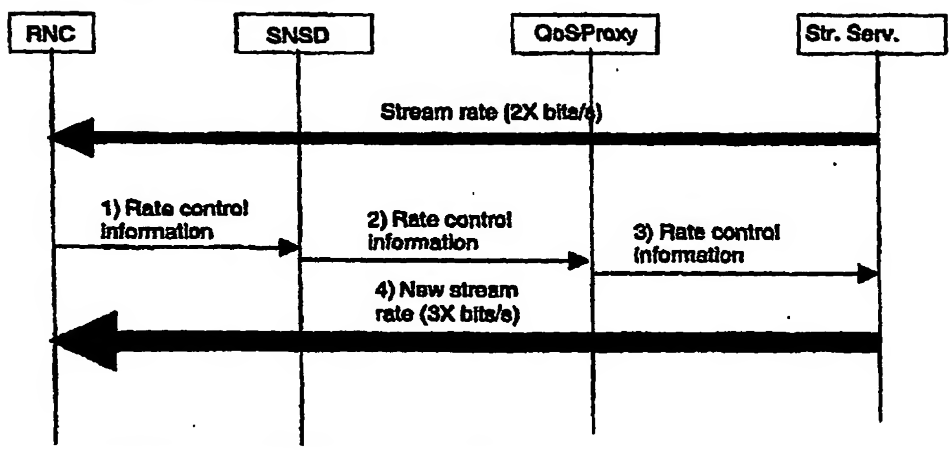

Figure 4 The Streaming Server streams with the rate of X bits/s.

1) The Radio Network Controlling node (e.g. an RNC or a BSC) informs the SNSD for a particular session how much the bit rate may be increased.

2) The SNSD either temporarily stores the information retrieved from the RNC or directly relays it further to the QoS-Proxy. (How this is employed may be decided during the configuration or the session set-up phase)

Read and Understood by:.. Date:.............

Read and Understood by:.. Date:.............

Uppgjord (även faktaansvarig om annan) - Prepared (also subject responsible if other)		Nr - No.		
EAB/LA/B J Petersson / H Emilion		EAB/QSI -2002:073 Uen		
Dokansv/Godk - Doc respons/Approved	Kontr - Checked	Datum - Date	Rev	File
EAB/QSIC (K. Boestad)		10/18/02	A	

3) The Rate Control information is relayed to the Streaming Server. E.g., by an RTCP report.

4) The Streaming Server starts to stream at the new rate (2X bits/s).

3.6 BENEFITS

Below exclusive benefits for this solution are given:

- The solution is very general. That is, it is not only applicable for Video services but also for TCP and MMS based services.
- The proposed solution is applicable to all types of applications. That is, the solution makes the applications network 'agnostic', meaning that the developers may put their focus on tailoring the applications for the mobile environment.
- It optimally balances the offered traffic over the air-interface with the back-end for all kind of person-to-content services over the packet switched domain leading to a better utilisation of the scarce radio resources.

Read and Understood by:.. Date:.............

Read and Understood by:.. Date:.............

+46 8 7520937

Uppgjord (även faktaansvarig om annan) - Prepared (also subject responsible if other)		Nr - No.		
EAB/LA/B J Petersson / H Emilion		EAB/QSI -2002:073 Uen		
Dokansv/Godk - Doc respons/Approved	Kontr - Checked	Datum - Date	Rev	File
EAB/QSIC (K. Boestad)		10/18/02	A	

CLAIMS

The present invention is not necessarily confined to the afore-mentioned described embodiments, but is instead defined by the following claims:

1) A system for optimizing any kind of data bit transfer over a wireless network, characterized in that it:
- gathers transmission quality information on assigned and currently employed air-interface bandwith for a bit transfer session
- inputs said transmission quality information as network feedback to a dedicated receiver of said information
- optimizes the session in accordance with certain algorithms so as to render improved quality of service to the user initiating the session.

2) A system as defined in 1) characterized in that
- said receiver of said network feedback is an application invoked by the user.

3) A system as defined in 1) characterized in that
- said receiver of said network feedback is a QoS proxy.

4) A proxy as defined in 3) characterized in that
- it is agnostic to being deployed in the service network or in the core network.

5) An apparatus for handling TCP transmission that involves radio transmission, characterized in that it:
- uses gathered radio transmission quality information as network feedback to actively adapt the TCP parameters so as to match the prevailing radio resource situation.
- prevents the triggering of the TCP congestion avoidance mechanism because of long air interface latency.

6) The apparatus according to 5) that further prevents the information downloading over TCP from taking too long during poor radio conditions characterized in that it
- uses said network feedback to instruct the application server to trim the transmitted-to-be information by trading "quality for time".

7) A method on handling TCP transmission involving radio transmission whereby a set of information is transmitted from a server to a client, the reception of said information to a network entity is acknowledged, and where said acknowledgment indicates the quality of the radio transmission as a kind of network feedback, characterized in that
– said network feedback is used by said network entity for adapting the TCP transmission parameters in order to prevent poor radio resource conditions from triggering the TCP congestion avoidance mechanism.

8) A method according to 7 characterized in that
- said network entity is a QoS proxy.

Read and Understood by:.. Date:............

Read and Understood by:.. Date:............

Uppgjord (även faktaansvarig om annan) - Prepared (also subject responsible if other)	Kontr - Checked	Nr - No.		
EAB/LA/B J Petersson / H Emilion		EAB/QSI -2002:073 Uen		
Dokansv/Godk - Doc respons/Approved		**Datum - Date**	**Rev**	**File**
EAB/QSIC (K. Boestad)		10/18/02	A	

9) A method according to 7 characterized in that
- said network entity is identical to said server.

10) A method according to 7 characterized in that
- said network feedback is also used for indicating to said server whether or not said set of information needs to be trimmed depending on said radio resouce situation, prior to commencing transmission of said set of information.

11) A system that allows for a general use of network feedback, defined as gathered information on radio transmission quality, for any application needing said network feedback, characterized in that it:
- contains an entity for knowing whether or not network feedback is required for a specific application or group of applications,
- contains a receiver, eg. an application server or QoS proxy, for selectively using network feedback with the purpose of adapting the session transmission rate to prevailing radio resource situation.

12) A method on efficiently using radio resources during person-to-content information transfer,
- whereby a session is set up between a client and a server, and
- information on radio transmission quality is gathered as network feedback,
- the browsing/downloading of said information takes place,
- during which the information transfer may be trimmed in real-time to fit certain radio resource constraints as measured by said network feedback, characterized in that
- network feedback is provided selectively.

13) A method according to 12) characterized in that
- prior to commencing the browsing/download, the information may be trimmed to fit certain radio constraints.

14) A method according to 12) characterized in that:
-the provisioning of said network feedback is done selectively, depending on the utilized application's need for said network feedback.

Read and Understood by:.. Date:.............

Read and Understood by:.. Date:.............

0203104-5

ERICSSON

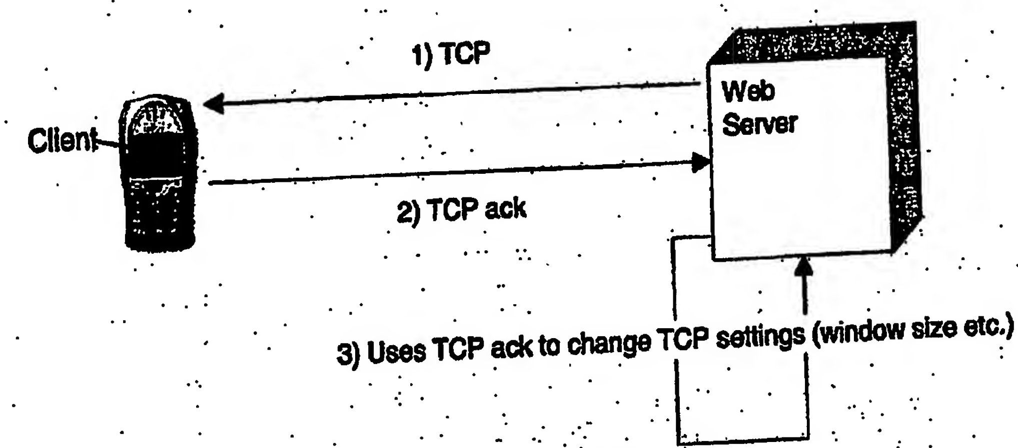

Figure 1A. Prior art: When a client is browsing or downloading information from a server, the client acknowledges the reception of data indicating the quality of the reception. The server uses this information to adapt transmission parameters (e g window or segment size) to the transmission conditions over the TCP link.

Problem: When the transmission link includes a radio connection there is a risk that bad radio connection conditions including many re-transmissions are misinterpreted by the server as Congestion, thereby triggering the TCP congestion avoidance mechanism (as the TCP transmission mechanism was not designed for radio transmission). Also, the radio conditions may change fast, but the feedback from the client comes comparatively late, which further reduces the server's ability to adequately react to changed radio transmission conditions.

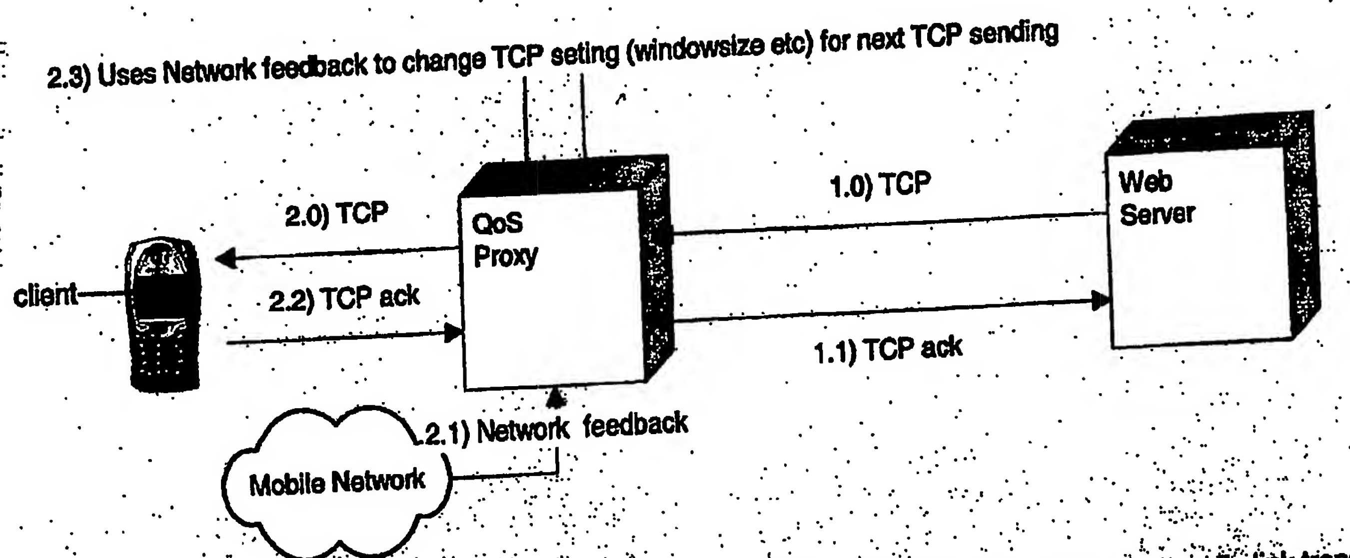

Figure 2A. An embodiment of the invention. Network feedback data regarding the quality of the radio link transmission is used to modify the TCP parameter settings (window and segment size etc). The network feedback may also be used to decide how to acknowledge reception to the Web server.

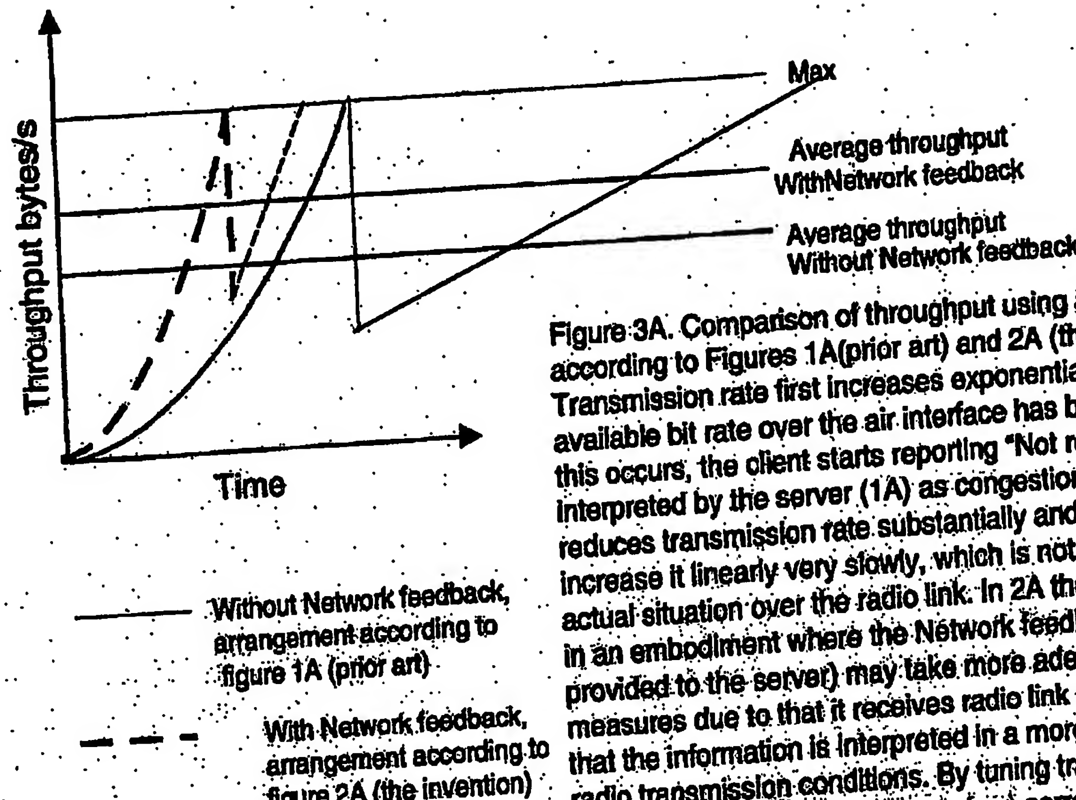

Figure 3A. Comparison of throughput using arrangement according to Figures 1A(prior art) and 2A (the invention). Transmission rate first increases exponentially, until max available bit rate over the air interface has been reached. When this occurs, the client starts reporting "Not received", which is interpreted by the server (1A) as congestion, whereby the server reduces transmission rate substantially and then starts to increase it linearly very slowly, which is not connected to the actual situation over the radio link. In 2A the proxy (or the server, in an embodiment where the Network feedback is directly provided to the server) may take more adequate and faster measures due to that it receives radio link information earlier and that the information is interpreted in a more accurate manner for radio transmission conditions. By tuning transmission parameters (e g TCP window and segment size) correctly, in may even be possible to adjust transmission rate so that it is long term stable.

This Page is Inserted by IFW Indexing and Scanning Operations and is not part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☐ FADED TEXT OR DRAWING
- ☑ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☐ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☑ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☐ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: ______________________

IMAGES ARE BEST AVAILABLE COPY.
As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.